MATE-T580: Quiz 3

Name:

Question 1

Here are two dataframes showing the same data of the stock price of a certain company over a 6 year period:

First 6 rows of dataframe `X`:

```
## # A tibble: 6 x 5
##    Year    Q1    Q2    Q3    Q4
##   <int> <dbl> <dbl> <dbl> <dbl>
## 1  2000    56    60    58    70
## 2  2001    78    72    80    93
## 3  2002    63    60    69    80
## 4  2003   101   111   108   103
## 5  2004   104   103   114   121
## 6  2005   130   135   134   150
```

First 6 rows of dataframe `Y`:

```
## # A tibble: 6 x 3
##    Year     Q Price
##   <int> <dbl> <dbl>
## 1  2000     1    56
## 2  2000     2    60
## 3  2000     3    58
## 4  2000     4    70
## 5  2001     1    78
## 6  2001     2    72
```

Select the line of code below that transforms `X` into `Y`:

A.

```
Y <- spread(X, Q, Price, Q1:Q4) %>% mutate(Q = parse_number(Q))  %>% arrange(Year, Q)
```

B.

```
Y <- gather(X, Q, Price, Q1:Q4) %>% mutate(Q = parse_number(Q)) %>% arrange(Year, Q)
```

C.

```
Y <- spread(X, Q, Price) %>% mutate(Q = parse_number(Q))%>% arrange(Year, Q)
```

D.

```
Y <- gather(X, Q, Price) %>% mutate(Q = parse_number(Q))%>% arrange(Year, Q)
```

Question 2

For the same dataframes `X` and `Y`, Select the line of code below that transforms `Y` into `X`:

A.

```
X <- spread(Y, Q, Price, Q1:Q4) %>% arrange(Year)
```

B.

```
X <- gather(Y, Q, Price, Q1:Q4) %>% arrange(Year)
```

C.

```
X <- spread(Y, Q, Price, sep="") %>% arrange(Year)
```

D.

```
X <- gather(Y, Q, Price, sep="") %>% arrange(Year)
```

Question 3

When joining two dataframes `M` and `N` by some key (e.g customer id), which is a valid reason to use `left_join(M, N, by=c("id"="id"))`:

A.

Retaining the subset of observations where the key matches across both dataframes `M` and `N`

B.

Retaining all observations from the two dataframes, whether the key matches or not

C.

Retaining all observations where the key is in `M` and not caring about `N`

D.

Retaining those observations where the key is in `M` but is not in `N`

Question 4

Which of the following statements represents the ***best*** motivation to investigate missing data in a dataset:

A.

Certain machine learning algorithms will produce an error if the input dataframe contains missing data

B.

Imputing missing data will improve the performance of any machine learning algorithm

C.

Imputing missing data makes the code run more efficiently

D.

It is prudent for a Data Scientist to understand why data is missing in order to come up with an appropriate imputation strategy

Question 5

Which two functions can be used to explore the structure of a dataframe? Write your answer below: